Danmarks Tekniske Universitet

Skriftlig prøve, den *23. maj 2012* Side 1 af 10 sider

Kursus navn: *Introduktion til Bioinformatik*

Kursus nummer: *27611*

Hjælpemidler: *alle*

Varighed: *4* timer

"Vægtning: *Angivet ved de individuelle opgaver*

# 27611 Eksamen Sommer 2012

Dette sæt indeholder 6 opgaver (side 1-10) – check at du har alle 10 sider.

Opgave 1 – Parvis alignment (10%)

Opgave 2 – UniProt (25%)

Opgave 3 – Proteinstruktur (20%)

Opgave 4 – BLAST (20%)

Opgave 5 – Multiple alignment og fylogeni (15%)

Opgave 6 – Binding af peptider (10%)

En online version af opgavesættet vil være tilgængeligt fra kursets lektionsplan under selve eksamen (Onsdag 23. maj 2012 klokken 9:00-13:00). DNA/Protein sekvenser og Accession-koder kan kopieres direkte herfra - det er *ikke* meningen at sekvenserne skal tastes ind i hånden.

Lektionsplan: http://wiki.bio.dtu.dk/teaching/index.php/27611:_Kursusplan_for_for%C3%A5r_2012

Svar til opgavesættet skal skrives enten i rå tekst (fx i JEdit) eller i et tekstbehandlingprogram såsom Microsoft Word. **Gyldige formater er .txt, .doc, .docx og .rtf**. I Opgave 5 skal der afleveres et billede – det kan enten indsættes i et Word dokument eller uploades separat som .png, .jpeg eller .gif.

Svaret skal uploades på CampusNet under kursus 27611 (under "Opgaver → Sommereksamen 2011"). **Husk at gemme seneste version af dokumentet inden du uploader svaret.** Når du afleverer får du en kode som skal skrives i feltet "Afleveringskode" nedenfor.

**VIGTIGT:** Dit studienummer skal fremgå af filnavnet (fx. s022717.doc eller s022717.txt) og skal også stå i starten af dokumentet (fx: "Studienummer: s022717")

Udfyld denne forside og aflever den til eksamensvagten.

Navn: _____

Studienummer: _____

**Afleveringskode: _____**

# Ang. brug af Internettet

## Trådløst internet:

Du skal koble dig på det helt normale DTU Wireless system, ligesom til øvelserne.

## Online materialer:

Linksamlingen til bioinformatik serverne er her:
http://wiki.bio.dtu.dk/teaching/index.php/Linksamling_for_27611

## BEMÆRK:

- I er ikke begrænset til kun de links der findes her – det er tilladt at søge information andetsteds.
- Det er **IKKE** tilladt at kommunikere med andre over nettet under eksamen.
- Der vil blive taget stikprøver af netværkstrafikken for at sikre dette.

## Hvad gør man hvis en web-server ikke virker:

1) Verificer at input-data er i korrekt format. Forkert inputdata er i næsten alle tilfælde årsagen til problemet.
2) Skift browser (f.eks. Firefox eller Chrome i stedet for Internet Explorer eller Safari). Visse webservere har problemer med visse browsere.
3) Prøv evt. at finde en alternativ server med samme funktion (Google).
4) Rapporter fejlen til eksamensvagten - den kursusansvarlige vil så blive tilkaldt.

## HUSK altid:

*”Don’t panic”*

Held og lykke med eksamen.

- Anders, Morten, Thomas, Bent og Henrik

## Opgave 1: Parvis alignment (10%)

Betragt de to nedenstående alignments:

```
   PAIRWISEALIGN        PAIRWISEALIGN
X: || |.||.    |  og Y: || |.||  |  |
   PA-RVISL----N        PA-RVIS--L--N
```

Og de to scoringssystemer
A: Matrix: BLOSUM62, Gap opening: -5, Gap extension: -5
og
B: Matrix: BLOSUM62, Gap opening: -10, Gap extension: -1

BLOSUM62 matricen er givet nederst på denne side. Husk at angive mellemregninger i spørgsmål a) og b), ikke bare resultatet.

a) Udregn (manuelt) alignment scores for X og Y givet scoringssystem A, hvor alle gap-positioner koster lige meget. Hvilket af de to alignments er optimalt?
b) Udregn (manuelt) alignment scores for X og Y givet scoringssystem B, hvor der er forskel på positionerne i hvert gap. Hvilket af de to alignments er optimalt?
c) Hvilket scoringssystem tror du er mest realistisk? Begrund dit svar!

BLOSUM62:

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 1 | 6 | -3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | -1 | -4 | -3 | -3 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | -1 | -2 | -1 | -2 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | -2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | -2 | -2 | -3 | -3 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | -2 | -2 | 2 | -3 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | -1 | -3 | -2 | -2 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | -2 | -1 | -1 | -1 | -1 | 1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | 1 | -3 | -2 | -2 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | -2 | -2 | 0 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | 2 | -3 |
| Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

## Opgave 2: UniProt (25%)

I 1994 publicerede tidsskriftet *Molecular Membrane Biology* et review med titlen "Cleavable signal peptides are rarely found in bacterial cytoplasmic membrane proteins" – se abstract her: http://www.ncbi.nlm.nih.gov/pubmed/8019598. Kort fortalt var budskabet, at mens transmembranproteiner i eukaryoter ofte har signalpeptider, så er det sjældent tilfældet i bakteriers plasmamembran. Du skal nu prøve at bruge UniProt for at se, om artiklens påstand holder.

a) Find først det protein, forfatterne beskriver som "*exceptional*" i abstractet, nemlig "*major coat protein of bacteriophage M13*" Hvilken søgestreng skal man bruge for at finde det så præcist som muligt?
b) Hvad er proteinets UniProt ID (entry name), dets anbefalede navn, og navnet på genet der koder for det? (Hvis du ikke kunne løse spørgsmål a) så får du UniProt accession her: **P69541**)
c) Som forfatterne siger, har dette protein både et signalpeptid og et transmembran-domæne. Hvilke positioner omfatter de? Indsæt aminosyresekvenserne for de to regioner i FASTA-format i din besvarelse.
d) Brug nu krydsreferencen til GenBank i UniProt (og proteinets gen-navn) til at finde den *DNA-sekvens*, der koder for proteinet. Indsæt DNA-sekvensen i FASTA-format i din besvarelse.
e) Hvor mange proteiner i UniProt er annoterede som transmembranproteiner ifølge "Subcellular location" feltet? (Tænk foreløbig ikke på om der er eksperimentel evidens). Skriv både antallet og den søgestreng du brugte. (**Hint:** se øvelsen i "Forudsigelsesmetoder")
f) Hvor mange af proteinerne i spørgsmål e) er fra hhv. eukaryoter og bakterier? Skriv både antallene og de søgestrenge du brugte.
g) Hvor mange af proteinerne i spørgsmål f) har signalpeptider? Skriv både antallene (for hhv. eukaryoter og bakterier) og de søgestrenge du brugte.
h) Angiv antallene i spørgsmål g) som procenter af antallene i spørgsmål f) og svar derudfra på om artiklens påstand holder – er der oftere signalpeptider i eukaryote transmembranproteiner end i bakterielle?
i) Svar på spørgsmål h) igen, men nu med den modifikation at der skal være eksperimentel evidens for annoteringerne (både for at de er transmembranproteiner og for deres signalpeptider). Holder artiklens påstand, undersøgt på denne måde? Skriv dine mellemregninger og søgestrenge.
j) Gør til sidst rede for hvad vi har ignoreret i denne undersøgelse (**Hint:** Hvilke *subcellular locations* kan bakterielle transmembranproteiner have? Sammenhold det med hvad der står i abstractet til artiklen).

## Opgave 3: Proteinstruktur (20%)

I denne opgave skal du se nærmere på proteiner af cadherinfamilien. CHDR3 fra mennesker har følgende sekvens:

```
>CDHR3
MQEAIILLALLGAMSGGEALHLILLPATGNVAENSPPGTSVHKFSVKLSASLSPVIPGFP
QIVNSNPLTEAFRVNWLSGTYFEVVTTGMEQLDFETGPNIFDLQIYVKDEVGVTDLQVLT
VQVTDVNEPPQFQGNLAEGLHLYIVERANPGFIYQVEAFDPEDTSRNIPLSYFLISPPKS
FRMSANGTLFSTTELDFEAGHRSFHLIVEVRDSGGLKASTELQVNIVNLNDEVPRFTSPT
RVYTVLEELSPGTIVANITAEDPDDEGFPSHLLYSITTVSKYFMINQLTGTIQVAQRIDR
DAGELRQNPTISLEVLVKDRPYGGQENRIQITFIVEDVNDNPATCQKFTFSIMVPERTAK
GTLLLDLNKFCFDDDSEAPNNRFNFTMPSGVGSGSRFLQDPAGSGKIVLIGDLDYENPSN
LAAGNKYTVIIQVQDVAPPYYKNNVYVYILTSPENEFPLIFDRPSYVFDVSERRPARTRV
GQVRATDKDLPQSSLLYSISTGGASLQYPNVFWINPKTGELQLVTKVDCETTPIYILRIQ
ATNNEDTSSVTVTVNILEENDEKPICTPNSYFLALPVDLKVGTNIQNFKLTCTDLDSSPR
SFRYSIGPGNVNNHFTFSPNAGSNVTRLLLTSRFDYAGGFDKIWDYKLLVYVTDDNLMSD
RKKAEALVETGTVTLSIKVIPHPTTIITTTPRPRVTYQVLRKNVYSPSAWYVPFVITLGS
ILLLGLLVYLVVLLAKAIHRHCPCKTGKNKEPLTKKGETKTAERDVVVETIQMNTIFDGE
AIDPVTGETYEFNSKTGARKWKDPLTQMPKWKESSHQGAAPRRVTAGEGMGSLRSANWEE
DELSGKAWAEDAGLGSRNEGGKLGNPKNRNPAFMNRAYPKPHPGK
```

Sekvensen kan også findes på dette link: http://wiki.bio.dtu.dk/teaching/images/2/2a/CDHR3.fasta

1. Find repræsentative strukturer for CDHR3 i PDB (www.pdb.org).
    a. Hvad er PDB-ID for de første fem hits? [**Hint**: En PDB-ID har formen 1XYZ].
    b. Hvilken af de fem første strukturer har den højeste tekniske kvalitet? [**Hint**: Ramachandran-plot kan findes under ”Geometry”-fanen på hver enkelt strukturs PDB-side]. Begrund dit svar.

2. Sammenlign de to strukturer 1EDH og 3K5S (fragmenter af hhv. E-cadherin fra mus og T-cadherin fra høns) i PyMOL. Læs begge strukturer ind i PyMOL og overlejr (engelsk: align) dem med følgende kommando:

    ```
    align 3K5S, 1EDH and chain A and resi 1-100
    ```

    a. Hvilken aminosyre i 3K5S svarer strukturelt til rest 59 i 1EDH? Angiv restnummer og aminosyretype for begge strukturer. [**Hint:** du kan få sekvenserne at se i PyMOL ved Display → Sequence i menuen].
    b. Er de fysisk-kemiske egenskaber bevaret ved denne aminosyresubstitution? Begrund dit svar.
    c. Hvor mange punktmutationer (nukleotidsubstitioner) skal der bruges for at ændre aminosyrerest 59 i 1EDH til den tilsvarende rest i 3K5S? [**Hint:** du kan f.eks. finde den genetiske kode her: http://en.wikipedia.org/wiki/DNA_codon_table].

# Opgave 4: BLAST (20%)

## Del 1: Ukendte gener

Efter din deltagelse i kurset Introduktion til Bioinformatik er du nysgerrig efter, hvad du kan bruge al den nye viden til i praktis. Du har derfor øjne og ører åbne for at se, hvor muligheden byder sig. En dag hører du en kollega tale om tre DNA sekvenser, som han desværre har mistet navnene på, og dermed ikke kender oprindelsen af. I din iver efter at bruge dine nye kundskaber, tilbyder du din hjælp med at identificere sekvenserne.

I dit svar til kollegaen SKAL du skrive hvilke værktøjer og databaser du vælger at bruge (samt hvorfor du vælger at bruge dem, så han kan forstå dine overvejelser). Dine svar skal også dokumenteres med relevant information.

De tre sekvenser ligger på følgende link:
http://wiki.bio.dtu.dk/teaching/images/8/85/3sekvenser2012.fasta

a) Hvilke organismer kommer de tre sekvenser fra?

b) Hvilke(t) gen(er) er der tale om?

c) Hvilket enzym/protein koder genet/generne for?

## Del 2 - Find proteinsekvensen

Din kollega er meget imponeret over din tilgang til at løse hans problem og håber derfor du kan hjælpe ham yderligere. Han kunne godt tænke sig at få den oversatte protein-sekvens fra alle tre organismer, hvilket du hurtigt siger at det er noget der kan løses på flere måder.

d) Beskriv kort minimum *to måder* hvorpå man kan komme frem til protein-sekvensen.

e) Indsæt protein-sekvensen i FASTA format for alle tre organismer i dit svar. Find selv på passende navne så din kollega kan se forskel på de tre sekvenser (undlad at bruge danske bogstaver, dvs æ, ø og å).

## Del 3 - Homologi

Efter at have fundet ud af at den ene sekvens kommer fra *Homo sapiens* kommer din kollega i tanke om at der er en af de to resterende sekvenser der er mere beslægtet med mennesker end den anden. Da det kun er disse to tættest beslægtede sekvenser han ønsker at arbejde med tilbyder du derfor igen din hjælp. Igen skal du veldokumentere dine svar med relevant information (brugte værktøjer/databaser, Signifikans værdier, Navn/ID på det mest signifikante match)

f) Hvis du sammenligner sekvenser mellem forskellige organismer, på hvilket niveau (DNA eller protein) vil du så foretrække at gøre dette (hvor er signalet stærkest)?. Begrund dit svar!

g) Brug BLAST og vælg den database du mener vil virke bedst, og undersøg om der findes et signifikant match (for begge sekvenser der ikke er humane) til en human sekvens. Hvilken sekvens har det bedste match til den humane sekvens? Giver det mening? Begrund dine valg og dokumenter dine resultater.

## Opgave 5: Multiple alignment og fylogeni (15%)

a) Den følgende liste af UniProt-accessionkoder svarer til det samme protein fra en række forskellige organismer. Konstruer en FASTA-fil som indeholder proteinsekvenserne, og indsæt den i din besvarelse. Hvilken funktion har dette protein?

```
P00918
P00919
P00920
P00921
P00922
P27139
G7MZP3
P07630
Q92051
B5X3I8
O04846
```

b) Lav et multiple alignment af proteinsekvenserne og indsæt det i din besvarelse.

c) Benyt UniProt-siden for det humane protein (uniprot-accession P00918) til at finde hvilke tre residues (aminosyrer) der indgår i det aktive site. Angiv deres positioner i den humane sekvens, og hvilke residues der er tale om.

d) Find de tre tilsvarende positioner i det multiple alignment (**tip:** brug evt. Jalview). Angiv deres positioner *i alignmentet*. Angiv for hver af disse tre positioner, hvor konserverede de er (hvor mange sekvenser har samme aminosyre som mennesket på den pågældende position?).

e) På basis af dit svar til sidste delspørgsmål, hvad synes du så om alignmentet? Er der noget der kunne gøres anderledes / bedre?

f) Lav et fylogenetisk træ på basis af alignmentet fra b). Placer roden ved at bruge planten (*Arabidopsis* — O04846) som outgroup. Hvilken art er ifølge dette træ nærmest beslægtet med *Mus musculus* (P00920)? Inkluder et plot af det rodfæstede (rootede) træ i din besvarelse.

## Opgave 6: Binding af peptider (10%)

Du har i dit arbejdet fundet 14 peptider der binder til en given receptor

**AENLWTTVY**
**AENLWVTPY**
**AENLWVTTY**
**FENLWVTVY**
**VENLWVTVY**
**AENLWVTVY**
**NENLWVTVY**
**TENLWVTVY**
**YENLWVTVY**
**AENPWVTVY**
**DELVDPINY**
**FPFKYAAAF**
**WEFCQPILL**
**SEAIHTFQY**

a) Beskriv ud fra disse data hvilke 2 positioner, der med størst sandsynlighed er mest vigtige for binding af peptider til den givne receptor. Beskriv kort hvordan du er kommet frem til dit svar.

**Hint:** Kig på informations-indholdet på de forskellige peptid-positioner

b) Du arbejder på at finde nye peptider, der kan binde til den givne receptor og har fået adgang til et sæt peptider fra din konkurrent. Uheldigvis er du kommet at blande peptiderne sammen med to peptider fra et andet studie. Kan du ud fra de data der var opgivet i spørgsmål a) finde ud af hvilke 2 af nedenstående 4 peptider, der med stor sandsynlighed vil binde til din receptor? Beskriv også her kort hvordan du er kommet frem til dit svar

**1) AENLWVHGM**
**2) AENLWVHGS**
**3) WEFLQPILL**
**4) SEAIHTFQY**

c) Det viser sig nu, at alle peptider med aminosyren “E” på position 2 er patenteret af din konkurrent. Du har derfor valget mellem at erstatte “E” på position med 2 med “A” (alanin, som er en lille og hyppigt brugt aminosyre i naturen) eller “D” (asparaginsyre, som er en ladet aminosyre). Hvilken af de to aminosyrer vil du vælge at indsætte på positive 2 i stedet for “E” således at effekten på bindingen til din receptor er mindst mulig? Begrund dit svar.#